LAB 13

STAT 28

April 27, 2017

Welcome to the lab 13! You will implement the regression and classification trees in this lab.

This will be a very short lab. Feel free to work on the two projects and ask questions if you have time left.

Regression tress with the diamond dataset

Read the data.

```r
diamonds <- read.csv("diamonds.csv")
diamonds <- diamonds[sample(1:nrow(diamonds), 1000), ]
head(diamonds)
```

```
##       carat     cut color clarity depth table price length.in.mm
## 5968   0.90    Good     F     SI1  57.6    60  3950         6.34
## 33809  0.32   Ideal     E    VVS2  62.3    56   842         4.37
## 40400  0.34   Ideal     D    VVS1  61.7    57  1133         4.49
## 17670  1.59 Premium     F     SI2  62.2    58  7123         7.49
## 21249  1.80   Ideal     H     SI1  62.2    57  9399         7.79
## 8882   0.90    Good     G    VVS2  62.6    63  4485         6.10
##       width.of.mm depth.in.mm
## 5968         6.37        3.66
## 33809        4.40        2.73
## 40400        4.52        2.78
## 17670        7.45        4.64
## 21249        7.71        4.82
## 8882         6.14        3.83
```

Exercise 1

(a) Grow a tree using the function `rpart` (The arguments are the same as `ls`) using the default `cp` (0.01).

```r
library(rpart)
set.seed(20172828)
# insert your code here
# diamond.tree <-
```

(b) Plot the tree using the `rpart.plot` function in package `rpart.plot`.

```r
# if "rpart.plot" package is not installed on your machine, run the following code to install
# install.packages("rpart.plot")
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 3.3.2
```

```r
# insert your code here
```

(c) What is your predicted price for the diamond with the following infomation.

- carat: 1.01
- cut: Good
- color: F

- clarity: S11
- depth: 63.9
- table: 59
- length.in.mm: 6.29
- width.of.mm: 6.32
- depth.in.mm: 4.03

```
My price prediction is ...
```

Classification regression - Customer Retension

This is a customer retension dataset from the last lab. The data set includes information about:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

Read the data.

```r
retension <- read.csv("customer_retension.csv", stringsAsFactors = FALSE)
retension$SeniorCitizen[retension$SeniorCitizen == 0] = "Yes"
retension$SeniorCitizen[retension$SeniorCitizen == 1] = "No"
retension = retension[retension$MultipleLines != "No phone service", ]
retension = retension[retension$OnlineSecurity != "No internet service", ]
retension$customerID = NULL
retension$Churn[retension$Churn == "Yes"] = TRUE
retension$Churn[retension$Churn == "No"] = FALSE
retension$PhoneService = NULL
retension$PaymentMethod = as.factor(retension$PaymentMethod)

test.set = sample(nrow(retension), 500)
retension.test = retension[test.set, ]
retension = retension[-test.set, ]
```

Exercise 2

(a) Grow a full tree using the function rpart (The arguments are the same as ls). HINT: For the rpart function, the default value of cp is 0.01. By setting cp to a negative number, the tree will fully grown.

```r
# insert your code here to fit a tree.
# retension.fit <-
```

(b) Plot the full tree using the rpart.plot function in package rpart.plot.

```r
# insert your code here to plot a full tree
```

(c) It is usually not a good idea to grow a full tree. Full tree can easily result in over-fitting. Now use printcp on your output object from rpart. Choose the cp with the smallest cross-validation error, use it to grow a tree and plot the tree you get.

```r
set.seed(20170427)
# # insert you code here to select the tree with the smallest CV error.

# best.cp <-
# best.tree <-
# # plot the tree
```

(d) Use function `predict` on the test data set `retension.test`. Use `type = "class"` to get the predited class. (Optional) Calculate the accuracy of the tree model. How does it compare with the logistic regression you implemented in lab 12?

```
# # insert your code here for prediction
# prediction <-
# # insert your code here to calculate the accuraray
# accurary <-
```